Detection of Non-Small Cell Lung Cancer adenocarcinoma using supervised learning algorithms applied to metabolomic profiles

Rondon-Soto, Diego[1] [0000-0002-1326-8909] Vela-Anton, Paulo Camilo[1] [0000-0001-5454-1118]

[1]Peruvian University Cayetano Heredia, Department of Biomedical Informatics

Abstract. Lung cancer is the most frequent and mortal of all types of cancer for both genders. Approximately 80% of the newly diagnosed lung cancers are non-small cell lung cancers. Early diagnosis improves the chances of survival. Machine learning allows us to process a considerable number of variables involved in this disease. Using metabolites as attributes for the analysis, we can discern lung cancer patients from healthy patients. In addition, machine learning algorithms reveal us which metabolites has a determining contribution in the classification. The objective of this study is to demonstrate the accuracy, sensitivity and specificity of a supervised learning algorithm to classify and predict non-small cell lung cancer, using concentration values found in the serum and plasma metabolome of afflicted and healthy humans. We obtained the dataset from the Metabolomics Workbench repository, which contains 335 samples and 139 known metabolites detected. Of all the models applied, Random Forest Classifier obtained the highest accuracy. It can classify participants according to diagnosis with > 75% accuracy in serum samples. Important serum metabolites for the classification included aspartic acid, fructose, and tocopherol alpha. Cystine, pyruvic acid and tocopherol alpha for plasma. The specified metabolites are strongly associated with this condition, and are potential biomarkers for the disease. By giving clues for an earlier diagnosis, this study remarkably contributes in the field of personalized medicine, and the appreciation of the biological processes of lung cancer.

Keywords: Lung Cancer · Metabolomics · Machine Learning.

1 Introduction

Lung cancer is the most common type of cancer in the world. During the 2018, it had the highest indicators for incidence and mortality (11.6% and 18.4% respectively) for both genders [1]. However, Non-Small Cell Lung Cancer (NSCLC), a type of epithelial lung cancer, is more dangerous and frequent in males than females [1]. For the year 2040, it is expected an increase in the incidence of 71.38% for both genders [1].

There are several methods for the diagnosis of lung cancer, such as X-rays. If this test does not clearly reveal an abnormal mass or nodule, clinicians may do an additional computed tomography (CT) [2]. Sputum cytology helps to detect tumor cells in the

phlegm [3]. Biopsy is the gold-standard test for the detection of different types of cancers [4-6]. Depending on the method of tissue extraction, physicians may do a bronchoscopy, the introduction of tubular camera through the trachea into the lungs [7], or a mediatinoscopy, which needs a surgical incision at the base of the neck to extract sample of the lymph nodes [8]. Another method is needle biopsy, which consist of a needle through the chest cavity into the lungs guided by a clinician [9]. These diagnostic methods are effective, however, they require expensive instruments and specialized personnel.

Metabolomics is the science that identifies and quantifies the total metabolites on the cell and extracellular environment. The metabolites are substances produced during the metabolism that give a functional reading of the physiological state of the body [10]. Consequently, we can elucidate what is happening in the organism along the tumorigenesis.

The human metabolome is influenced by exogenous factors such as diet, drugs, physical activity and other environmental aspects [11, 12], and by endogenous factors such as age, gender, IBM and disease [13, 14]. Therefore, a metabolomics' approach is an efficient way for the early detection of certain types of diseases like osteoarthritis [15] and gestational diabetes [16]. Many studies about several types of cancer apply metabolomics to find potential biomarkers [17].

Computer science advances at hectic pace in our times; computers facilitate the collection and interpretation of data and results. During the last decades, computational tools have improved their skills to learn and predict. In the field of artificial intelligence, machine learning (ML) allows the computer to "learn" given patterns and then be able to recognize them.

In this study, we test the capacity of prediction of different models of ML: Decision tree classifier (DTC), support vector machine (SVM), random forest classifier (RFC), K-nearest neighbors (KNN), and logistic regression (LR) [18, 19, 20, 21]. Our purpose is to train an algorithm to discern lung cancer patients from healthy patients using their metabolomic profiles.

2 Data Collection

We collect the data from the Metabolomics Workbench Repository, a free access metabolomic databank [22]. The identification code of the project in the databank is PR000293. ST000368 and ST000369 are two independent case-control studies that build the whole information of the project. Both studies investigated NSCLC adenocarcinoma by the untargeted metabolomics approach using gas chromatography time-of-flight mass spectrometry to analyze the metabolome of serum and plasma samples. We use the data obtained from the serum and plasma results. The analysis included only the known metabolites from both groups. We selected only 139 metabolites as attributes. The data consist of 166 samples, 92 with cancer diagnosis and 74 controls in the serum group. 169 samples divided in 95 with cancer diagnosis and 74 controls in the plasma group. Fig 1 shows the workflow to collect and preprocess the raw data obtained.

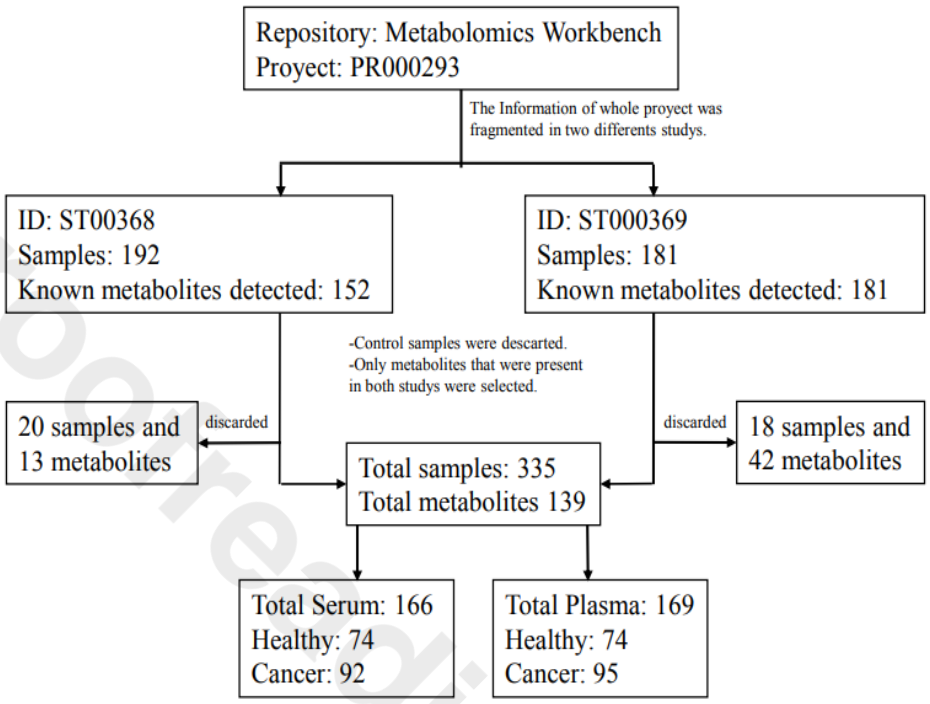

Fig. 1. Data extraction flow chart

3 Method

The data was preprocessed as follows: non-existent values were replaced by the median of the pertinent input variable. The entire data set was normalized and 60 cancer patients and 60 healthy patients were chosen from each dataset to train the algorithm. The remaining samples were used for validation. Since this is a binary classification, we applied several supervised learning algorithms to find which had the best performance. RFC is a supervised classifier that assembles prediction results of a number of classification and regression trees, and it is commonly used in high variance data. In parallel to RFC analysis, we performed LR, SVM, KNN and DTC on the same datasets. All of them were performed with packages from Orange (Version 3.21). Parameters used in each method and platform are shown in table 1.

Table 1. Parameters used for each method in Orange platform.

Method	Parameter
Random Forest	Number of trees: 150 Maximal number of considered features: 5 Fixed random seed: 42

	tree depth: unlimited Stop splitting nodes with maximum instances: 5
Logistic Regression	Regularization: Lasso (L1), C=1
SVM	SVM type: SVM, C=1.0, ε=0.1 Kernel: RBF, Numerical tolerance: 0.001, Iteration limit: 100
kNN	Number of neighbors: 5 Metric: Euclidean Weight: Uniform
Decision Tree	Pruning: at least two instances in leaves, at least five instances in internal nodes, maximum depth 100 Splitting: Stop splitting when majority reaches 95% (classification only) Binary trees: Yes
Ensemble bagging	DecisionTreeRegressor(), n_estimators=500, bootstrap=True, oob_score=True, random_state=1, base estimator = DecisionTreeRegressor

For all the methods, the classification performance was evaluated by 5-fold cross validation, where 80% of the data were used for feature selection and model construction, and the area under the Receiver Operation Characteristic curve (AUROC), Classification Accuracy, Precision, Recall and F1 Score of the model were evaluated based on the hold-off one fifth data. From the RFC we selected the features with the highest importance, representing the important metabolites associated with the diagnosis. For all the models, the 5-fold cross validation was replicated 300 times, to obtain a report of the average AUROC, Classification Accuracy, Precision, Recall and F1. We made a confusion matrix, and calculated true positive rate (TPR), true negative rate (TNR), kappa values and Matthews correlation coefficient (MCC) only for RFC.

4 Results

Table 2 presents values of area under the curve (AUC), Classification accuracy (CA), F1 score (F1), precision (Pr) and Recall (Re). The method with the best performance appears in bold. Table 3 displays relevant values after validation with the test sets.

Table 2. Values of AUC, CA, F1, Pr and Re by method in Orange 3.21 platform.

Method	**Serum**					**Plasma**				
	AUC	CA	F1	Pr	Re	AUC	CA	F1	Pr	Re
RFC	**0.919**	**0.807**	**0.798**	**0.797**	**0.807**	**0.638**	**0.584**	**0.576**	**0.577**	**0.584**
LR	0.804	0.771	0.776	0.782	0.771	0.555	0.538	0.540	0.542	0.538
SVM	0.841	0.753	0.727	0.725	0.759	0.544	0.541	0.525	0.528	0.541
KNN	0.845	0.759	0.753	0.749	0.759	0.553	0.550	0.550	0.549	0.550
DTC	0.723	0.765	0.762	0.760	0.765	0.563	0.571	0.571	0.571	0.571

Table 3. TPR, TNR, MCC and Kappa Value for Random Forest Classifier.

Measure	Random Forest Classifier	
	Serum	Plasma
TPR	0.65	0.70
TNR	0.67	0.43
Kappa Value	0.43	0.37
MCC	0.54	0.22

We obtained the highest values for CA and AUC by using the RFC algorithm (0.80 and 0.91 respectively), in comparison with other methods.

RFC was the best method to diagnose cancer; it is not affected by outliers and missing values. The model also provides us which metabolites are the most relevant to separate cancer patients from controls. Fig. 2 and Fig. 3 show the concentration intervals of the pertinent metabolites from serum and plasma samples. In respect to serum, fructose indicated low concentration in a greater number of profiles with NSCLC adenocarcinoma in comparison with healthy profiles. Aspartic acid and tocopherol alpha registered more samples of all concentration. In plasma, cystine demonstrated low concentrations in cancer cases and high concentrations in healthy subjects. Pyruvic acid presents medium to high concentrations in cancer patients. Tocopherol alpha has an alternating behavior in relation to the concentration intervals and the state of the individual.

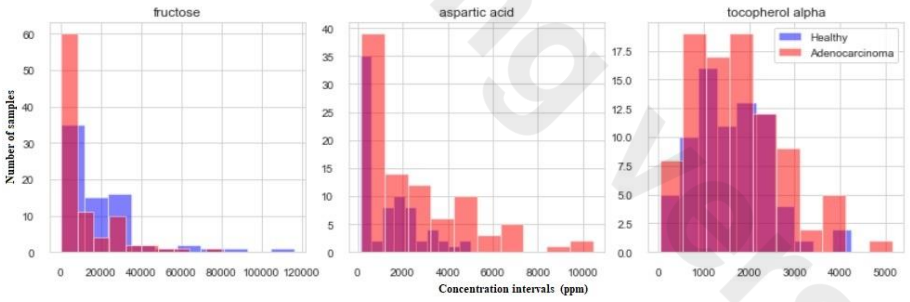

Fig. 2. Number of samples per concentration interval of the three principal metabolites found in the serum samples, fructose, aspartic acid and tocopherol alpha.

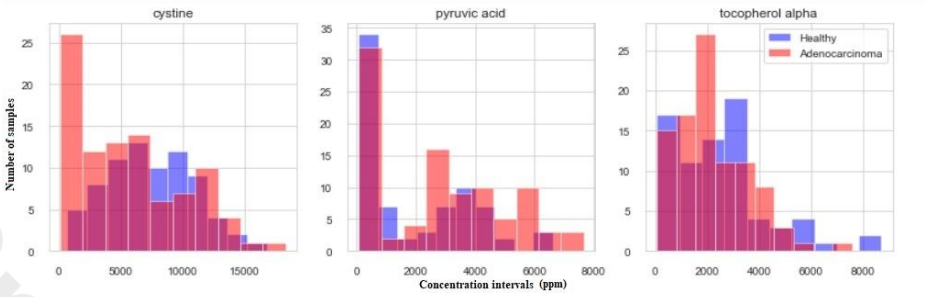

Fig. 3. Number of samples per concentration interval of the three principal metabolites found in the plasma samples, cysteine, pyruvic acid and tocopherol alpha.

5 Discussion

In this study, we went through the best model to diagnose lung cancer in two datasets with metabolomic profiles from different tissue, conformed by 139 attributes each one. RFC proved to be the best method to discern between individuals with cancer and healthy in serum and plasma. The precision to separate afflicted patients from controls was higher in the serum group.

The metabolites used for the classification differ between groups. In the case of serum samples, the algorithm found that fructose, tocopherol alpha, aspartic acid, acetophenone, beta alanine and citrulline were decisive components for the prediction. In the plasma group, the algorithm showed that cystine, pyruvic acid, alpha tocopherol, arabinose and lactamide are useful metabolites to detect NSCLC adenocarcinoma.

Metabolites used by the algorithm have a biological relationship with the development of cancer and its proliferation. Previous studies shown that the increase in concentration of aspartic acid, citrulline and β-alanine in serum are important characteristics of NSCLC [23]. Fructose promotes lung adenocarcinoma, cell survival and metastasis through GLUT5 [24]. The administration of cystine in tumors in vivo increases the use of glutamine by the cancer cells [25]. Nishith K. *et al* propose lactamide as a biomarker found in plasma for lung cancer [26].

The algorithm identified alpha-tocopherol as an important metabolite for detection of NSCLC adenocarcinoma. Several studies observed an association between low concentrations of alpha tocopherol and risk of lung cancer [27]. This metabolite acts as an anti-oxidant, preventing DNA damage by free radicals.

The results are significant for the treatments of patients due to its value in a better understanding of the metabolic pathways and the physiopathology of NSCLC adenocarcinoma. The metabolites found are potential biomarkers that could help in the early diagnosis of this cancer in order to improve the survival rate of cancer patients.

The first limitation we found was the collinearity of the data when training the models without preprocessing. An explanation for collinearity is that a metabolic pathway involves several metabolites. In this degree, the sub production or overproduction of one can affect the production of another, maintaining a direct proportionality between them [28].

Another drawback was the unbalanced datasets. It is common to find this characteristic in the biomedical field. It affected the recognition of patterns for the minority class, in this case, the healthy patients. It is necessary to increase the size of the datasets found in the metabolomics databases, in order to make a better analysis from a broader number of samples in this studies [29].

6 Conclusion

The RFC proved to be the errorless method to detect NSCLC adenocarcinoma when applied to the metabolomic profiles of serum and plasma. Detection accuracy was higher in serum (approximately 80%) compared to plasma (around 60%). We found that multiple metabolites are strongly associated with the detection of NSCLC adenocarcinoma. Fructose, alpha-tocopherol and aspartic acid were decisive attributes for detection in serum profiles, while cystine, pyruvic acid and alpha-tocopherol were relevant in plasma profiles. We recommend including variables such as age, sex and smokers condition to have a more reliable prediction.

7 References

1. Bray, F.: Global Cancer Statistics 2018. GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries 394–424 (2018).
2. Patil, SA.: Chest X-ray features extraction for lung cancer classification. Journal of Scientific and Industrial Research (India). 69(4), 271–7 (2010).
3. Ammanagi, A.: Sputum cytology in suspected cases of carcinoma of lung (Sputum cytology a poor man′s bronchoscopy!). Lung India 29(1), 19 (2012).
4. Salomon, L.: Prostate Biopsy in the staging of prostate cancer. Prostate Cancer Prostatic Dis 1(2):54–8 (1997).
5. Afyon, M.: Liver Biopsy is the Gold Standard at Present, How about Tomorrow?. Viral Hepatitis Journal 22(2), 67–8 (2016).
6. American, T.: Performance and Practice Guidelines for Excisional Breast Biopsy. American Society of Breast Surgeons 1–3 (2014).
7. Manthous, C.: Flexible bronchoscopy (Airway Endoscopy). American Journal of Respiratory and Critical Care Medicine 191(9), P7 (2015).
8. Hoeijmakers, F.: Mediastinoscopy for Staging of Non-Small Cell Lung Cancer: Surgical Performance in The Netherlands. The Annals of Thoracic Surgery 107(4), 1024–31 (2019).
9. Chojniak, R.: Computed tomography-guided transthoracic needle biopsy of pulmonary nodules. Radiología Brasileira 44(3), 99–106 (2007).
10. Roessner, U.: What is metabolomics all about?. Biotechniques 46(5), 363–5 (2009).
11. Menni, C.: Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. Metabolomics 9(2), 506–14 (2013).
12. Floegel, A.: Linking diet, physical activity, cardiorespiratory fitness and obesity to serum metabolite networks: Findings from a population-based study. International Journal of Obesity 38(11), 1388–96 (2014).
13. Auro, K.: A metabolic view on menopause and ageing. Nature Communications 5, 1–11 (2014).

14. Kochhar, S.: Probing gender-specific metabolism differences in humans by nuclear magnetic resonance-based metabonomics. Analytical Biochemistry 352(2), 274–81 (2006).
15. de Sousa, EB.: Metabolomics as a promising tool for early osteoarthritis diagnosis. Brazilian Journal of Medical and Biological Research 50(11), 1–7 (2017).
16. Mao, X.: Metabolomics in gestational diabetes. Clinica Chimica Acta 475 ,116–27 (2017).
17. Palmnas, MSA.: The future of NMR Metabolomics in cancer therapy: Towards personalizing treatment and developing targeted drugs?. Metabolites 3(2), 373–96 (2013).
18. Scholkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. 1st edn. MIT Press, Cambridge, MA, USA (2001).
19. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 4th edn. Morgan Kaufmann, USA (2016).
20. Cuperlovic-Culf M: Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling. Metabolites (2018).
21. McCullough, B: On the accuracy of linear regression routines in some data mining packages. WIREs Data Mining and Knowledge Discovery (2019)
22. Sud, M.: Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic acids research *44*(D1), D463-D470 (2015).
23. Klupczynska, A : Evaluation of serum amino acid profiles utility in non-small cell lung cancer detection in Polish population. Lung Cancer 100:71-76 (2016).
24. Yuanyuan, W.: Fructose fuels lung adenocarcinoma through GLUT5. Cell Death and Disease 9:557 (2018).
25. Alexander, M.: Environmental cystine drives glutamine anaplerosis and sensitizes cancer cells to glutaminase inhibition. eLife 7; 6: e27713 (2017).
26. Nishith K.: Serum and Plasma Metabolomic Biomarkers for Lung Cancer. Bioinformation 13(6): 202–208 (2017).
27. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. Ann Epidemiol (1):1-10 (1994).
28. Fani, R : Origin and evolution of metabolic pathways. Physics of Life Reviews. 6(1):23-52 (2009).
29. Calabrese, F.: Are There New Biomarkers in Tissue and Liquid Biopsies for the Early Detection of Non-Small Cell Lung Cancer? Journal of Clinical Medicine 8, 414 (2019).